Heuristic Approach to Organic Chemistry

Martin Schatz
The University of Texas at Austin

Abstract

An important question in organic chemistry is how does a chemist synthesize chemicals in the most efficient way beginning with chemicals readily available. These problems are known as synthesis problems. Currently, methods based on databases (Chen, 2006) and methods related to quantum mechanics (Jorgensen, 1998) are used frequently to solve this problem. When solving synthesis problems by hand, a key step is to identify the functional groups of molecules as this gives the experimenter a set of applicable reactions. The functional groups required for different reactions are experimentally derived and accepted into a knowledge base. Instead of relying on this knowledge base for functional group identification, perhaps there is a way to gain the same level of accuracy for functional group identification with large datasets of reactions. This paper investigates a heuristics method of determining the functional groups of a certain reaction type.

1 Introduction

The field of chemistry attempts to understand interactions in the natural world occurring at the microscopic scale. One of the main concerns of chemists is how to accurately predict the outcomes of chemical reactions. Organic chemistry is a division of chemistry that abstracts away many of the fine details of the physical quantities. This division deals with synthesis problems more than other divisions and more readily accepts rules for reactions that do not have as thorough as a physical basis. Because many of the physical details are not dealt with as much in organic chemistry, more abstract rules can be generated.

This lends many problems in organic chemistry to be more easily expressed as rule based systems than problems in other divisions of chemistry. Two commonly used approaches to solving the problem of reaction prediction are discussed in this paper: using computers to simulate the reaction as it progresses (Jorgensen, 1998) and predicting the outcome of reactions based on rules derived from experimentation (Chen and Baldi, 2009).

The former benefits from the fact that simulations can be based on quantum mechanics (Repasky et al., 2002) which can very accurately describe the system, however it suffers from large computational costs to achieve that level of accuracy. Scientific computing demands for resources have increased dramatically (Con, 2007), resulting in computing centers devoting large amounts of resources to these problems. Simulating reaction systems are particularly computationally complex especially when high levels of accuracy are desired. Currently, accurate and precise predictions can be made if the simulations are based on quantum mechanics. The trade-off for accuracy and precision is the resource requirements.

If high levels of precision are not desired, as prediction problems in organic chemistry often are not, database approaches can be used to solve these problems. Instead of solving mathematical equations that describe the physical system, general rules for reactions can be stored and applied to inputs to determine the molecular outputs of the reaction (Chen and Baldi, 2009). This solution to prediction problems is more applicable to organic chemistry as it, as stated earlier, is more accepting of rules with smaller amounts of physical bases. The downside to database/rule based approaches is that the database must be filled with definitions of reactions and can become massive due to the multitude of exceptions that occur. Compounding this is that each output of one reaction can possibly be the input to another reaction. Thus solving prediction problems can often

lead to search spaces too large to searched by brute force (Todd, 2005).

Both of the described methods, in my opinion, lack a certain quality that makes the intuition behind the methods difficult to see or reason about. Even though the above methods solve the problem of prediction to a high degree of accuracy, the methods themselves seem more related to calculators than to thought processes. Originally, machine learning was considered as a different method of solving the problem of prediction. The hope was to have a system that was able to get good levels of accuracy for solving prediction problems (most certainly not the same levels as the above mentioned methods but still good enough for organic chemists) while having a system that modeled a kind of thought process for solving the problem. A desire of the system was to also have lessened the amount of resources (time and memory) needed to gain good levels of accuracy.

Due to time constraints and failed methods, the larger problem of prediction was abandoned, but instead the smaller problem of functional group identification was investigated. This paper presents a heuristics method for solving prediction problems which hopes to lay the groundwork for future work to solve the larger problem of synthesis while still having an intuitive methodology. As mentioned earlier, functional group identification is a key problem to solve when presented with a prediction problem and simulations are not allowed.

This project uses many simple ideas in computer science but applies them in a different context in the hopes that they can more easily express ideas.

2 Problem

2.1 Chemical Reactions

Reactions in organic chemistry rely on a number of characteristics of the chemical system, but generally speaking the reactions rely on the chemical properties of the molecules. The properties are defined by the structure and composition of the individual molecules. Different elements can have vastly different chemical properties. By replacing a single atom with an atom of a different element, the molecule can cease to be bonded together, or the molecule can have completely different chemical properties. Different chemical properties can also come about by changing the bonding structure of the molecule.

Of course, not all molecules are chemically active, and similarly not all parts of a molecule need be chemically active. Organic chemists, through numerous experiments, have defined a set of chemically active subgroups of molecules referred to as functional groups. Functional groups are the keys to synthesis and reaction prediction work. If the functional groups of the molecules can be identified, the outcome can consequently be predicted (to a first approximation).

For clarity and brevity, a molecule used to perform the reaction is referred to as a *reactant*, and a molecule created by the reaction is referred to as a *product*. Chemically active groups of a reactant are referred to as the *functional groups* of the reactant, and groups of a product which are formed will be referred to as the *inverse functional groups* of the product.

2.2 Problem of Synthesis

A synthesis problem in chemistry is a problem involving products that are desired and some starting materials (the reactants). The problem then is to devise a method comprised of a set of reactions that will transform the reactants into the desired products (Vollhardt and Schore, 2007).

Variations on this problem can be made. For instance, instead of trying to specify a method that will transform the reactants into the products, try to specify the products that are created by applying a reaction method to the reactants. This variation can be thought of as a prediction problem where the question asks what the predicted products of a reaction are.

More challenging prediction problems can be specified as well, such as by only specifying the starting materials and having the chemist use their chemical knowledge to predict the outcome of the reaction that is known to exist yet is not specified. Figure 1 graphically depicts the different variations of synthesis problems. In this figure, the desired information is designated with a question mark and the known information for the problem is stated. Figure 1 also depicts which step(s) in the reaction process going from reactants to products is/are missing from the problem.

Most organic chemistry courses are taught by introducing a functional group and explaining different types of reactions they can be involved in and why the reaction types occur. If the student is able to remember the functional groups required for the reaction type to take place and also remembers how the reaction type transforms the reactants, these challenging prediction problems can be solved readily.

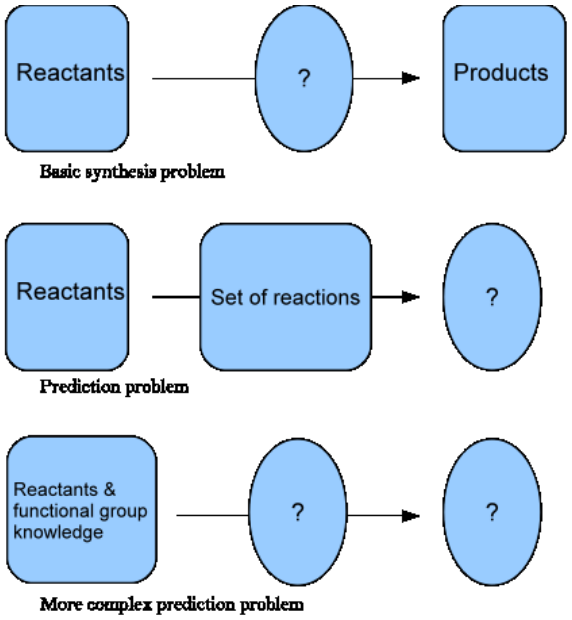

Figure 1: Variations of synthesis problems

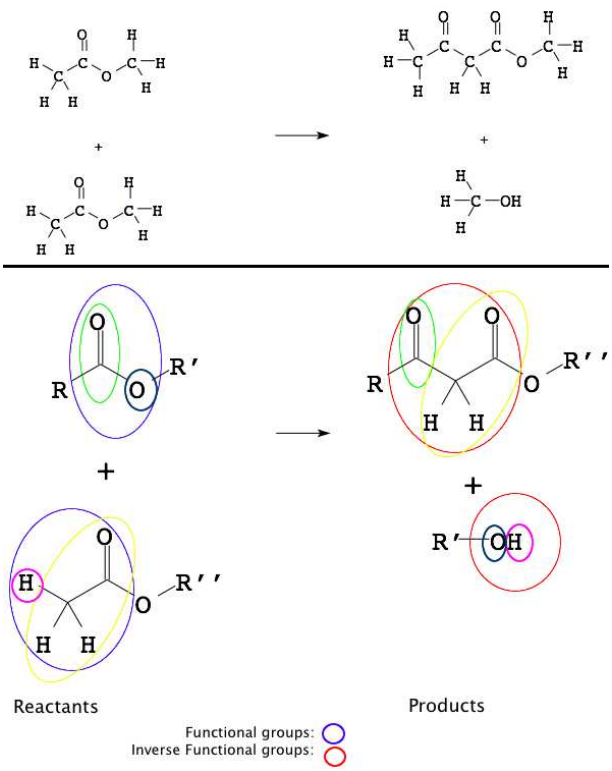

Figure 2: Reaction transformation of functional groups. As the reaction proceeds, the functional groups can attach to different sections of molecules involved in the reaction. Here, colored groups in the product are similarly colored groups found in the reactants. Non-colored groups do not change bonding at all.

By identifying the functional groups of the reactants in a given problem, the student can use the learned knowledge to correctly predict the outcome of the reaction. Thus, identifying the functional groups of the reactants is a critical step in correctly predicting the outcome of these prediction problems.

2.3 Project Problem

Given a set of reactions belonging to the same type of reaction and their corresponding results, can a set of heuristics accurately determine the functional and inverse functional groups of the reaction type?

2.3.1 Motivation

If a simple set of heuristics is able to identify both the functional and inverse functional groups of the reaction then it is believed that enough information is available to estimate the correct transformation of reactants that occurs in the reaction.

The argument that identifying these groups is enough to accurately estimate how the reaction type proceeds relies on the meanings of "functional" and "inverse functional" groups. Since the functional groups are the chemically active portions of the molecule, they are the groups that cause change to occur in the reaction and are not spectators of the reaction. Inverse functional groups are the groups created by the reaction by interactions within the functional groups. If both of these sets of groups can be accurately identified, they will contain all the information associated with how the reaction proceeded.

Most organic reactions are simply a set of bond transformations within the set of reactants. Because of this, the set of inverse functional groups and set of functional groups should contain the same atoms. By examining these sets and fitting functional groups onto the inverse functional groups, a hypothesis can be made as to how the reaction proceeded. It is this reasoning that leads to the belief that by identifying the functional and inverse functional groups, the rules of the reaction can be estimated.

Figure 2 graphically depicts this idea. The functional groups and inverse functional groups are circled in blue and red respectively. Within each functional/inverse functional group, subgroups in the reactants are connected to a different set of atoms in the product. Whether the subgroup is found in the

set of reactants or products, it is identified by the same colored circle surrounding it. This shows how a functional group transforms into the inverse functional group as it goes through the reaction.

If the rules of the reaction can correctly be identified, this allows for the larger problem of reaction prediction to be solved. Thus, identification of not only the functional groups, but also the inverse functional groups is a key step to solving the larger problem of prediction.

2.3.2 Statement

Fundamentally though, the problem of functional group identification is the same as inverse functional group identification. The only difference is that in one case the reactants are being examined as opposed to the products. This paper more specifically focuses on identifying the inverse functional groups of the reaction, though this is a technicality as the inverse functional group can be considered a functional group of the reverse reaction.

To solve this problem computationally, a representation of molecules that can be understood by computers is needed. The most intuitive way to represent a molecule is as a graph with nodes that correspond to the atoms comprising the molecule and edges that correspond to the bonds connecting the molecule. Since bonds can have different strengths associated with them, weights are applied to the edge that correspond to the strength of the bond. For this problem, molecules are assumed to be uncharged entities and bonds are assumed to have integral bonding strengths.

2.3.3 Computational Constraints

Since this problem relies heavily on graph data structures, many applicable algorithms have large space or time complexities. One of the stated desires of the original problem was to alleviate some of the resources needed to compute the solutions to prediction problems. Keeping this in mind, it was decided that the resources needed to solve the smaller problem of functional group identification should be kept small.

A common problem of this project is subgraph isomorphism. Throughout the proposed method, comparisons between the graph representations of subgroups of molecules are made to update tables of probability information. To accurately do this, an algorithm that solves the problem of subgraph isomorphism is required. Even though it is not known whether to be NP-Complete, the number of subgroups compared makes efficiency vitally important to the method.

2.3.4 Examples

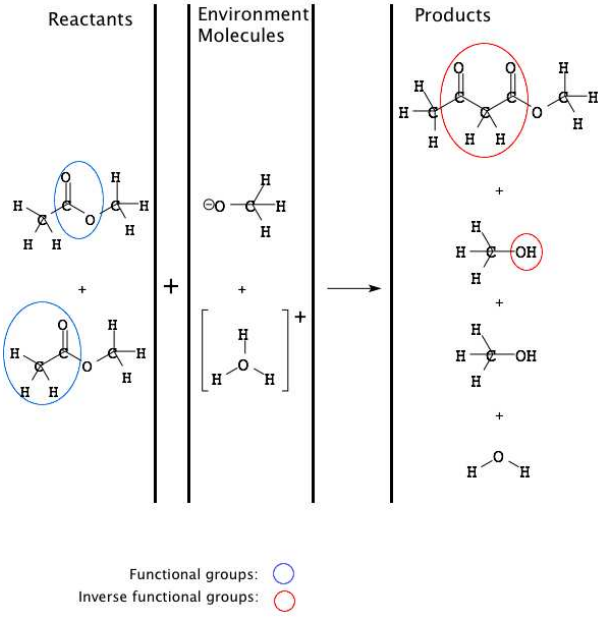

Figure 3: Claisen Condensation example

Figure 3 shows an example of a *reaction type* in organic chemistry known as a claisen condensation (Vollhardt and Schore, 2007). This figure outlines everything involved in the reaction while hiding the processes that occur in the reaction itself. The environment molecules are necessary for the reaction to take place, however they do not have lasting effects on the system. The bottom two products (methanol and water) are simply an equilibrium form of the environment molecules. Thus, though environment molecules are often needed to allow reactions to occur, they often do not cause a fundamental change in the system and therefore are not considered in this project.

As molecules become larger, their graphical representations become cluttered. Hydrogen and carbon are two of the most commonly seen atoms in organic chemistry, so a shorthand is created to minimize the clutter of the drawings.

Carbon atoms most often defines the backbone of the organic structure. The structure of the organic molecule is desired to be kept, but it can be assumed that the structure is defined by carbon atoms unless otherwise specified. A shorthand then is to omit the labeling of atoms at the vertices of two lines if the

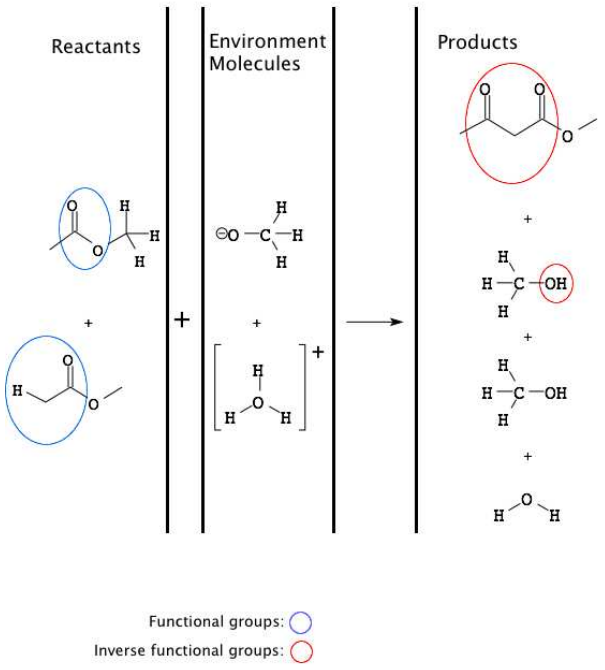

Figure 4: Claisen Condensation example shorthand

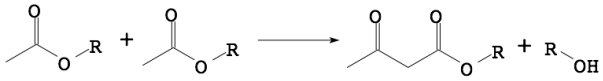

Figure 5: Claisen Condensation example shorthand

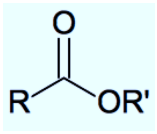

Figure 6: Ester functional group

atom is a carbon whose only purpose is structure. The lines are bent as they are due to an early theory of bonding devised by Linus Pauling (Anslyn and Dougherty, 2006).

Hydrogen atoms usually play no role in the reactivity of molecules. Their purpose is usually to saturate the various atoms to fill the empty orbitals of atoms (Vollhardt and Schore, 2007). If the orbitals are not filled, the atoms can become highly reactive causing the molecule to not behave as if it were isolated. A shorthand then is to assume that all carbons are saturated with hydrogens unless otherwise specified (ie. the bond is a non C-H bond). This removes all hydrogens from the description connected to carbons unless they are important for understanding the reaction.

Figure 4 shows the same reaction as Figure 3, except with shorthands applied.

The shorthands described are useful for individual reactions. However, many reactions belong to the same reaction type. Since a reaction type is generally only concerned with the transformations occurring with the functional groups of the reactants, a shorthand is used to represent the segments of the molecule which do not impact the reaction. Instead of fully describing the structures of the segments, the entire segment is replaced with an 'R'. Segments that are different in structure and also non-essential are differentiated from other segments by adding tick marks to the R in their description. Figure 5 shows the definition of the claisen condensation reaction type in this format (Vollhardt and Schore, 2007).

Esters, the functional groups involved in claisen condensation, comprise most of the reactant molecules' structure. Figure 6 shows the graphical description of an ester. The only new notation is the introduction of R groups. These are groups which are not crucial to the chemical activity of the functional group and can be essentially anything so long as the non-R structures exist. To state that two R groups need not necessarily be the same, quotation marks are used to differentiate between different R groups within the same functional group. In Figure 6, R' (attached to an oxygen molecule) may not be structurally equivalent to R (attached to the carbon atom).

3 Previous Work

There are currently a few well accepted methods to predicting the results of reactions in organic chemistry. The two methods (sets of rules for plausible

reactions and simulation based on quantum mechanical models of the system) mentioned earlier are predominantly used methods in chemistry (Chen, 2006).

3.1 Databases

As mentioned previously, reactions in organic chemistry can be predicted quite accurately using experimentally determined rules. These rules tell the chemist in general terms the requirements and steps needed to have the reaction take place. The more rules an organic chemist knows, the more plausible pathways to synthesizing the end product they can create. Most of these rules can be expressed with three pieces of information:

1. Required structures for reaction to take place
2. Extra environment information (solvent, temperature requirements, etc.)
3. Result of the reaction. This can be specified with either or both of the following:
    - How the transformation should proceed
    - Resulting structure of the transformation

The first item tells the chemist what combinations of molecules are needed for the reaction to take place at all. Without these required structures, the reaction will not take place.

The second item informs the chemist of environmental requirements for the reaction to occur. Some reactions require very low acidity and if this criteria is not met, the reaction will not occur, or will occur, but with differing results.

The third item informs the organic chemist of the general output of the reaction which is crucial for synthesis. As stated, there are two ways the chemist can deduce the output of the reaction. If the chemist is provided with how the reaction transforms the important molecules in the system, the products can be predicted. If this information is supplied, performing the deductions is no longer required. It is more important to know how the transformations proceed rather than the output as more information about the system is available to the chemist to make decisions.

If the three pieces of information are given for each reaction, numerous paths for the problem of synthesis can be created. Generally, the more rules a chemist knows, the stronger and more flexible the chemist is for solving synthesis problems. Unfortunately for chemists, the number of rules for synthesis has increased dramatically. It is now to the point that chemists cannot remember all the possible reaction rules and be able to quickly recall/use the information. Even if a chemist were able to remember all the reaction rules, it is unlikely an individual would be able to produce the plethora of plausible results given a set of reactants. With the advent of computers, this burden can be shifted from the chemist to the machine.

The rules of valid reactions can be encoded and stored in a database. Given enough processing power, a chemist can obtain the result of a synthesis reaction through a query of this database. Given the query, the machine can create a working set of possible products by applying each of the applicable rules. If any of these products match the requested product, the path can be returned to the chemist. Rules can be continually applied until all plausible rules have been exhausted. By doing this, the database method can return the set of reactions needed to create the product.

Similarly the problem of prediction can be solved. Since the database stores the transformation rules for different reactions, if the product of a reaction is requested and the database has the rules for the reaction in question, the database can apply the rules and return the product.

With the database approach to synthesis, as long as the database knows enough rules and sufficient processing resources are provided, most variations of the synthesis problem can be solved.

3.1.1 Analysis of Database Approach

The major advantage of a database approach for the storage of the rules governing reactions and the application of rules to queries approach for synthesis problems is that it takes a very large burden off the chemist allowing the chemist to focus on other more important things. Also, the database can be viewed as a repository for all the accepted rules of reactions in organic chemistry with enough description so as to be easily understood and relearned by another chemist.

A downside to using a database of organic chemistry reactions is that its size can be massive. In light of the database's massive size, it still may be unable to predict the outcomes of new reactions. A chemist, on the other hand, possesses the ability to think critically. This ability with the chemist's chemical intuition can be used to develop hypotheses concerning original reactions.

The database's massive size can be the cause of an additional problem as well. Most reaction pathways used to solve synthesis problems involve multiple reactions. This causes the products of one reaction to be the reactants of another. Since the number of steps taken to reach the products is not specified and each intermediate step can be viewed as a reaction problem, the number of possible pathways increases extremely quickly. This can cause database methods to apply heuristics to limit the number of steps allowed (Todd, 2005).

In the simplest implementation, a database approach is unable to apply the same critical thinking of a chemist as it lacks a notion of chemical intuition. Since the example is a never before seen reaction, the database is unable to make good predictions.

3.2 Quantum Mechanics

A milestone in the natural sciences was the development of quantum mechanics. The way that scientists thought about the world was completely rewritten with this new theory that better described the smallest scales of the universe. Through systems of equations, quantum mechanics is able to describe all observable physical quantities of particles and systems.

By applying operators and solving the equations associated with systems, scientists are able to obtain results that very accurately match experimental values (Repasky et al., 2002). Thus a system can be evolved over time to give the final state of the system. A reaction can therefore be simulated very accurately, allowing chemists to see the results of reactions with great confidence in the solution. Unfortunately, setting up the equations and, furthermore, solving them quickly becomes too difficult to be done by hand.

A set of libraries can be created to solve equations related to simulation of reactions. With this library and a fairly large amount of processing power, a computer is able to simulate a reaction as it would occur in a system. This allows the chemist to simply run a job describing the reaction through a computer and analyze the results. The problem of synthesis can be solved with these forms of simulations. The chemist can input the starting material into the simulator and view the time evolution of the reactions. This allows the chemist to look for pathways and see if the end result is what was predicted.

3.2.1 Analysis of QM Simulation

The major benefit of solving the problem of synthesis using this method is that the results generated describe what happens in the natural world with a high degree of precision. Unfortunately, the equations associated with even simple systems are difficult to solve on a computer within a reasonable amount of time and a moderate amount of processing power. Most synthesis problems rely on fairly complex and large scale systems. This results in upwards of weeks of computation required to view the result of a synthesis reaction. Thus, this method is applicable when precision is far more important than processing resources.

Though this method of synthesis prediction suffers from computational complexity, it does have a fundamental advantage over the database method: it does not rely on a set of experimentally determined rules describing organic reactions. This method works by relying on rules that describe particles more generally. This lower level of abstraction allows for the prediction of unknown reactions.

4 Why investigate a new method?

As stated in Section 3.2 and Section 3.1, there exist methods of solving organic synthesis problems which are accurate and very precise. Each method described though suffers from an inherent weakness that I believe is an incorrect model of how a chemist would approach the problem and therefore should be reexamined.

The method described in Section 3.2 solves the problem of organic synthesis with simulation which requires using computationally complex functions. In essence, methods based on simulation solve systems of equations by evolving the reaction in time to compute the results. I believe that this method does not correctly model how a chemist would approach a synthesis problem. It is the case that the mathematical functions used to simulate the reaction are extremely precise, however the essence of a chemists' intuition about the problem is lost on the computer. The computer then becomes a fast and powerful calculator.

In the case of a simple database approach (Section 3.1), the computer is given a set of rules by the chemist. These rules describe what a reaction type requires to occur and how it should proceed. The pitfall of this method is that the computer only has information about the rules given to it. With this idea, the computer is only able to apply known

rules to the input, resulting in an inability to predict the outcomes of unseen reaction types. Moreover, the computer becomes a powerful pattern matcher. An organic chemist can become powerful at pattern matching as well, however the chemist is able to understand new reaction types whereas a rule based program is unable to.

The idea behind the original project was to investigate a method of solving organic synthesis problems that more appropriately models how the chemical intuition comes about in the chemist. Self placed constraints were applied to satisfy my desire to have a method of reaction prediction that is less computationally intensive than simulations but still accurate enough for organic chemists to trust.

5 Data

A large set of sample reactions belonging to the same reaction type is required for the method to determine the functional and inverse functional groups of reactions. The database was not designed to minimize the time to compute the functional groups of the reactions, but rather to represent a full range of molecules with interesting characteristics, and as such included many larger sized molecules.

Since this method relies on a structural analysis of molecules in each reaction, a full structural description of the molecules is required.

5.1 Dataset Parameters

This method was applied to a varying number of randomly chosen sample reactions from a generated database of approximately 1000 reactions to predict the correct inverse functional group of one of the product set's molecules in each reaction within the dataset.

5.2 Data Description

For a given reaction, a data file is associated with it that describes the reaction. In each data file, the format is described as follows:

Figure 7 gives the format of the input file for a given reaction. A full description of the structure is required for each molecule involved in the reaction. This description is given in Figure 8. It is important to note that the abbreviated form of element names are used (ie, for "Carbon", "C" is used).

Bonds between atoms are given in the bond section of the molecule description. Chemically, a bond can have varying strength. For the purposes of the synthesis problem being solved, these values are integral in the range [1,3]. Higher value bonding does occur, though usually between metals and not with organic molecules.

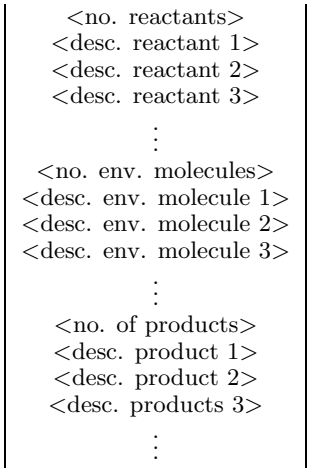

Figure 7: Data input format

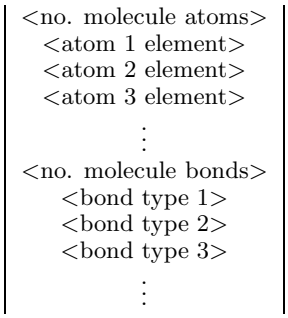

Figure 8: Molecule description

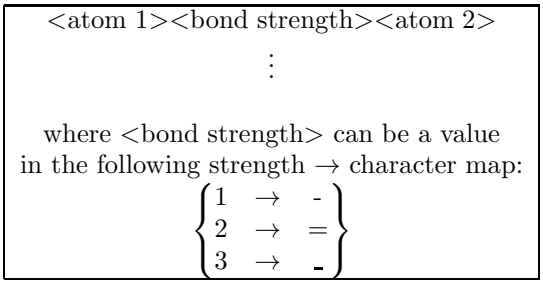

Figure 9: Bond description

When describing a bond, the connected atoms are given along with a character denoting the strength of the bond. Figure 9 shows the characters used to denote the strength of the bond. Each atom is labelled by the element name as well as an arbitrary index that is unique within the set of atoms of the same element. Atoms with different element types can share the same index since their label is ultimately different.

5.3 Data Source

Large databases of chemical reaction information were not available to access, so a program was created to perform, to a first approximation, the rules based synthesis routines needed for reaction prediction. Modifications were made to produce reactions in bulk and to generate the output in the format required for the machine learning method.

Since only one reaction type was investigated, the set of rules was not extensive and allowed for the generation of thousands of sample reactions of the same reaction type.

6 Approach

Multiple approaches were investigated before deciding on the method used. Each method provided advantages and disadvantages, but working in very strict constraints, many were rejected before the final approach was selected. Each approach implemented is described in the Appendix and explanation is given as to why they were or were not selected for functional group identification.

The methods investigated determining inverse functional groups of reactions belonging to the same reaction type. As stated earlier, this step is important to solving the greater problem of synthesis. Most methods implemented attempted to solve this problem deterministically, but the chosen method relies on a large data set and the use of heuristics to approximate the inverse functional group.

Here, the problem of finding the inverse functional groups in a reaction can be described as finding the "best" subgraph of a product such that it includes the most sites of transformation when the reactions occur. Defining "best" is somewhat subjective without the introduction of some chemical knowledge, but the basic idea is to have a tight fit around the chosen subgroup that shows the location of the transformation.

One of the desired traits of the method used was that it should model what a person with no chemical intuition considers when asked to solve the problem of identifying the inverse functional groups given a dataset of reactions. As this project is laying the basic foundation for better methods, a rudimentary one was investigated initially.

The method chosen attempts to mimic what a person might think of when given the task of identifying inverse functional groups without having any chemical background. For clarity, this method assumes that atoms belonging to inverse functional groups are inherently more important than the rest of the atoms of the molecule and that molecules contain regions of reactivity instead of being completely reactive.

6.1 Method

The problem of inverse functional group identification can be reduced to identifying which atoms are important and which are not. If each atom could be assigned a value based on its level of "importance", the correct atoms could be selected more easily. This leads to the introduction of some scoring function that defines the importance of each atom. Based on the relative scores, atoms with high "importance" scores could be considered as part of an inverse functional group. When first looking at a set of reactions, the more unique an atom is, the more it will stand out. The heuristics used to define the scoring function relate to the uniqueness of the atom in question.

6.1.1 Scoring function

Two measures of "importance" are made for each atom. One is based on the relative frequency of finding the same element within the molecule and the second is based on the uniqueness of the region surrounding the atom. The "importance" of each atom is then the sum of these two measurements.

A first idea for differentiating important atoms from others is to look at the element types. Elements seen more often than others might lead one to think they are somehow more important as they are seen less often in the molecule.

To represent this thinking, the proposed method assigns an initial score to the atom based on the relative frequency of the atom's type(what element the atom is). The following states more formally how this value is calculated:

A few functions need to be defined for calculating the score associated with only the atom element type.

$elem(a : Atom)$
$elem$ returns the type of atom a is.

$atoms(m : Molecule)$
$atoms$ returns a set containing all the atoms comprising the given molecule m.

$count(item : T, set : \{T\})$
$count$ returns the number of occurrences of $item$ in the given set.

Now the score associated with the atom a considering only its properties is defined as follows:

$$Molecule <: \{\}$$
$$E_m = \{elem(a) \text{ st. } a \ \epsilon \ atoms(m)\}$$
(duplicates allowed)
$$S_m(a) = \frac{count(elem(a), E_m)}{|E_m|}$$

Scores are represented in log space to avoid machine precision errors.

Similarly, bond strengths can vary in the molecule. This could be used as another source for assigning importance. An atom bonded by a stronger bond than normally seen could be viewed as more important than other atoms as has a more unique trait.

Extending this idea further, by looking at the atoms connected to the central atom can be examined for to differentiate two atoms of the same type in a molecule. This idea results in examining the different regions the atoms each reside in. The more unique a region looks, the more interesting the central atom may appear. Thus something related to the uniqueness of the location the atom in question resides in could be an indication of importance.

Larger regions will likely have a greater impact on the central atom as adding a single atom compounds the probability of finding the region further. This, however, seems counterintuitive as atoms far away from the central atom should have less of an effect on the central atom's uniqueness. To remedy this, some normalization of the scores based on regions surrounding the atom should occur. It may be difficult to see how much a region's uniqueness should impact the central atom when only considering a single molecule, so examining the regions of all molecules being considered may give a better idea as to the impact.

This idea is represented in the proposed method by calculating the relative frequencies of regions of increasing size surrounding the atom whose "importance" score is being calculated. Regions from all molecules being examined whose sizes match are considered.

Edit distances between graphs are being avoided as this is a costly operation for calculating the "importance" scores of each atom, each region is converted to a string that encapsulates all the items that compose the region but does not encapsulate the connectivity. Edit distances are avoided because earlier attempts to implement methods based on edit distances resulted in the constraints of the problem being broken.

Each score associated with a region is normalized by the size of the region defined by the number of atoms composing it. Since scores are represented in log space, this actually models an approximately exponential drop-off in score effect due to an increasing size of region. Regions of all size are considered in this part of the calculation for thoroughness. A more formal definition of this calculation is described below. First a few functions need to be defined:

$region(a : Atom, s : Int)$
region returns a graph composed of all the atoms at most s bonds away from a

$all\_regions(s : Int)$
all_regions returns a Map containing regions of size s from all molecules being examined as keys, and the count of each region as a value

$total\_count(map : Map <? \Rightarrow Int >)$
total_count returns the sum of all the values in *map*

The score associated with looking at regions surrounding the atom is calculated as follows:

$S_r(a) = 0$
for(size in $\{1 \ldots |atoms(m)|\}$):
  $c = all\_regions(size)[region(a, size)]$
  $reg\_sz = |atoms(region(a, size))|$
  $num\_reg = total\_count(all\_regions(reg\_sz))$
  $S_r(a) \mathrel{+}= \frac{1}{reg\_sz} * log\left(\frac{c}{num\_reg}\right)$

Unlike $S_r(a)$, $S_m(a)$ is already represented in log space. This was done to make the equation more readable. The score, S, assigned to the atom a is then the sum of $S_r(a)$ and $S_m(a)$.

$$S(a) = S_r(a) + S_m(a)$$

The proposed method only defines these two heuristics for calculating the score associated with each atom. After the scores have been calculated for all the atoms in a molecule, a threshold is applied to determine which atoms are of great enough importance to be considered part of the inverse functional group.

Once the scores have been assigned to each atom in the molecules, a simple threshold is applied to choose the atoms that should belong to the functional group. This threshold, T, is based on the mean score calculated and the relative spread of the scores.

Incorporating the deviation of the data allows for the removal of data effects such as having a large cluster of atoms with low "importance" scores affecting the data.

All atoms having an "importance" score lower than the threshold are estimated to be part of a functional group of the molecule. Since the scores are related to probabilities of different groups in the molecule and are in the log space, a higher "importance" score implies having a lower related probability which implies a more negative score. Therefore, atoms are considered part of a functional group if they are below the threshold. The threshold can be adjusted by setting a value f which tells the method how many standard deviations away from the mean the threshold should be set. A more formal description of the application of the use of the threshold for a given molecule is described below. First a few functions need to be defined:

$score(a : Atom)$
score returns the calculated "importance" score for a

$mean\_score$ = the average score of all atoms comprising the molecule m
std_score = the standard deviation of the scores of all atoms comprising m
f = a factor used to adjust the threshold
$T = mean\_score - (f * std\_score)$
the threshold of the program
$func\_atoms = \{\}$
(will eventually contain all atoms predicted to belong to a functional group)

The application of the threshold can now be defined as follows:

for(a in atoms(m)):
if($score(a) \leq T$)
$func\_atoms = func\_atoms \cup \{a\}$

At the end of this process, all the atoms predicted to belong to functional groups are in the set $func\_atoms$. Though this method implies that multiple functional groups can be found within a molecule which is allowed in chemistry. However, this problem is only concerned with finding one such group and thus the set of atoms connected to the most unique atom is considered the actual inverse functional group of the reaction.

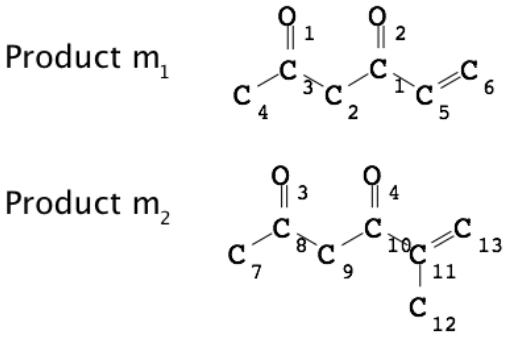

Figure 10: Sample product set

6.1.2 Example

Figure 10 shows a sample set of products belonging to separate reactions. The calculation of the "importance" score associated with one of the atoms will be calculated to show how the scoring function works. To make things simpler, hydrogens will be ignored in this example.

Firstly, the score for the atom C_1 will be calculated. The first step in assigning an "importance" score to C_1 is to calculate the relative frequency of carbon atoms in the product m_1.

$$\begin{aligned} E_m &= \{C, C, C, C, C, C, O, O\} \\ S_m(C_1) &= log\left(\frac{count(C, E_m)}{|E_m|}\right) \\ &= log\left(\frac{6}{8}\right) \\ &= -.12494 \end{aligned}$$

Now the score based on the region surrounding C_1 will be calculated. Unfortunately, this is more difficult to depict graphically, so a table will be created to show the calculation of $S_r(C_1)$

$size$	reg_size	c	num_reg	score term
1	4	4	5	$log(\frac{4}{5})/4$
2	6	1	2	$log(\frac{1}{2})/6$
3	8	2	2	$log(\frac{2}{2})/8$
4	8	4	4	$log(\frac{4}{4})/8$
5	8	8	8	$log(\frac{8}{8})/8$
6	8	8	8	$log(\frac{8}{8})/8$
7	8	8	8	$log(\frac{8}{8})/8$
8	8	8	8	$log(\frac{8}{8})/8$
			$S_r(C_1)$	-.07440

To help identify which groups are being selected

for each iteration, the following table was created:

size	set of atoms chosen for region
1	$\{C_1, C_2, O_2, C_5\}$
2	$\{C_1, C_2, O_2, C_5, C_6, C_3\}$
3	$\{C_1, C_2, O_2, C_5, C_6, C_3, C_4, O_1\}$
4	$\{C_1, C_2, O_2, C_5, C_6, C_3, C_4, O_1\}$
5	$\{C_1, C_2, O_2, C_5, C_6, C_3, C_4, O_1\}$
6	$\{C_1, C_2, O_2, C_5, C_6, C_3, C_4, O_1\}$
7	$\{C_1, C_2, O_2, C_5, C_6, C_3, C_4, O_1\}$
8	$\{C_1, C_2, O_2, C_5, C_6, C_3, C_4, O_1\}$

Now that both $S_r(C_1)$ and $S_m(C_1)$have been calculated, the final score assigned to atom C_1 is:

$$\begin{aligned} S(C_1) &= S_m(C_1) + S_r(C_1) \\ &= -.12494 + (-.07440) \\ &= -.19934 \end{aligned}$$

Test	Avg. ED	Avg. rxn. molecule ED	Actual func. group ED
		Baseline: all atoms selected	
9 consec. reactions	9.0	40	13
100 random reactions	9.88	36	13
200 random reactions	8.595	28.995	13
500 random reactions	9.29	28.678	13

Figure 11: Results

7 Results

The best metric for evaluating this method on determining the functional groups of molecules is to calculate the edit distance between the predicted functional group and the actual functional group. As stated earlier, this is an expensive process. Therefore, a few metrics were used to determine the quality of the method. Both a quantitative and qualitative metric were used. The quantitative metric gives a measurable result that can be used to compare different methods and the qualitative metric gives the scientist an idea of the shortcomings of the method.

7.1 Quantitative Metric

The quantitative metric compared the number of types of atoms and bonds in the actual functional group to the predicted functional group. Connectivity was not examined as that would result in an inexact subgraph matching problem.

An edit distance (ED) score is assigned to each of the predicted functional groups for the reactions in the dataset. The edit distance is calculated from the predicted inverse functional group to the known actual inverse functional group and is based on the count of each type of item comprising the actual functional group. Any discrepancy in the number of a certain type (be it bond type or atom type) is added to the edit distance for the predicted group. The final ED for the predicted functional group is the sum of the discrepancies for each type.

The tests run used 9 systematically chosen molecules and 100 randomly chosen molecules. The systematically chosen test was performed to see how the method reacts given a set of reactions that are all built upon one another. Tests with randomly chosen reactions were performed to show the method's performance under real circumstances.

The baseline mentioned in Figure 11 was chosen to give an upper bound on the maximum edit distance possible for the method. The baseline shows the average ED for the dataset if the method had chosen every atom within each molecule as part of the predicted functional group.

Based on the results summarized in Figure 11, the predicted functional group is off by at most the number of edits required to make the functional group. This metric suffers from not expressing how equivalent the structures of the functional groups are. If the method predicts an extra atom belonging to the functional group, the edit distance increases by two because an extra bond is needed to connect the atom. In this fashion, the quantitative metric can be misleading at times.

Based on the baseline chosen, the method's precision in choosing the correct functional group is low. Since the metric only shows the ED from the predicted functional group to the actual functional group, a predicted functional group could either be larger or smaller than the actual functional group and the result would be the same. With this information as well as the baseline values, the method appears to have large imprecisions in the prediction.

It appears that as the number of reactions increase, the method finds an optimum data set size. A max-

imum of 500 random reactions were tested because this was half the dataset. The result seems a bit worrisome in that it appears that more data can help find a better result, but too much data can have an adverse effect as well. One reason for this could be that the larger dataset increases the number of similar regions within the molecules, reducing the effect of the region based portion of the score. If this is true, it would be a good indication that the heuristics defined are not enough to model chemical intuition correctly.

Each test was run and timed to express some measure of the viability for using the method as a basis for solving the larger prediction problem. For 100 random reactions, the test took 3.5 minutes to finish and for 200 random reactions, 7 minutes were taken to complete, and 500 random reactions took 15 minutes to complete. Since the dataset was not designed to be biased towards the smaller sized molecules, but was designed to cover a range of molecular sizes, these timing results show that the proposed method does have promise when considering resource use. If better heuristics are implemented and the computation time stays near these, then I believe the goal of solving prediction problems using more representative methods of intuition is reachable.

The threshold chosen for accepting atoms as a part of a functional group was chosen for this project to be 0.8 standard deviations below the average. This was done prior to running a rigorous test to determine the optimal threshold.

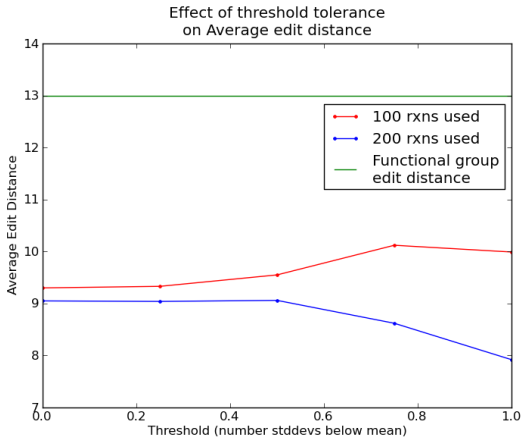

Figure 12: Threshold effect on Average ED

A test was performed to determine what the optimum number of standard deviations away from the mean the threshold should be set at to result in the most accurate results. Tests involving 100 and 200 reactions were performed while varying the threshold and the effect on the average edit distance was recorded.

The threshold values varied from 0-1.25. 1.25 standard deviations was chosen as a stopping point as this stated that only atoms with "importance" scores greater than 90% of the data would be included in the predicted functional group. Figure 12 summarizes the results of this test.

It has been stated that the quantitative metric does not express information regarding the connectivity of the predicted group. To get a better gauge for the structural similarity of the functional groups, the qualitative measurement is used. This is more useful for determining how appropriate the method is to the problem as it depicts structural differences.

7.2 Qualitative results

The quantitative metric compares two methods of functional group identification, however it does not express structurally how close the prediction method came determining the correct functional group. For that, images were created to see the predicted functional groups and compare them to the actual functional group.

Figure 13 and Figure 14 show examples of predicted inverse functional groups determined by the method outlined in this paper. To understand the images, the colors and labels need to be explained.

Each atom represents a node in the graphics and its filled color indicates what element its type is. An atom of type carbon is black, oxygen is red, and hydrogen is white. The label associated with each atom gives the atom's molecular label as well as its group number. The molecular label is used to distinguish between atoms of the same type, and the group number is used to show varying levels of "importance" of the atoms. The assignment of group numbers is not deterministic in nature which causes the colors of the predicted functional groups in the images to differ.

Bonds are depicted as edges in the images. The strength of the bond is depicted by the number of edges connecting the two atoms together. If the two connected atoms share the same group number, the bond shares the same group color. If they differ, then the color of the bond is black.

In these images, the functional group determined by the method connected by magenta bonds in Figure 13 and by yellow bonds in Figure 14.

Based on qualitative results, the proposed method still has problems isolating the correct single func-

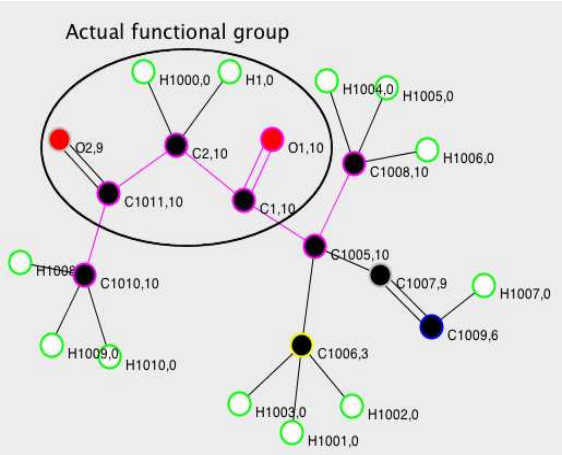

Figure 13: Actual functional group circled in molecule whose predicted functional group is magenta colored

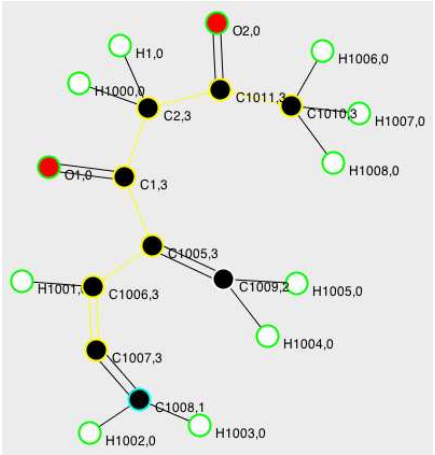

Figure 14: Sample molecule with predicted functional group highlighted in yellow

tional group. Figure 14 shows that the heuristics can fail by not selecting the oxygen molecules as part of the functional group.

A database approach to solving the problem of prediction was implemented to generate the dataset for this project. If this were modified instead to produce the inverse functional group of a given reaction, the results would show an exact matching to the actual functional group. This, however is misleading as the database already stores the functional groups of the reaction as well as the transformation that occurs. Therefore, the database simply needs to track the atoms through the reaction and then return the connected result.

The proposed method performs poorly in comparison as it must determine the inverse functional group without having knowledge of how the atoms in the reactants match to those in the product.

8 Future Work

This paper focuses on identification of the functional and inverse functional groups of the reaction type. A simple set of heuristics is used to solve this problem, but based on the results it still suffers from problems. A future step would be to investigate better ideas to incorporate to the heuristic model and compare results.

As stated earlier, this problem is crucial to solving prediction problems in organic chemistry. A next step could be to build on the accomplishments in this paper and examine how best to identify the transformation rules of a reaction type.

After this step is done, two approaches can be taken. Either solving larger synthesis problems can be explored, or different reaction types can be included in the dataset. By attempting the former, the original goal of the project can be accomplished, whereas the latter could involve investigating methods to estimate which reactions belong to the same reaction type.

9 Conclusion

Applying a simple heuristic to identify the functional groups of molecules is a first step in modeling how a person might initially attempt to solve the problem. Without any chemical information being provided, it is difficult to identify the functional groups of the reaction type by simply examining structures of the molecules involved.

The results of this project show that the simple heuristics applied to this problem are not adequate enough to correctly identify the functional group of a specific reaction type.

Even though the proposed method does not accurately predict the functional groups of the reaction type, it does model some instinctive thinking processes a person might go through when given the same problem.

10 Appendix

The following section describes all the methods implemented and rejected before choosing the final method for solving this problem.

The initial approach to solving this problem was to identify where the reactant molecules best fit on the products. If the largest groups were placed on top of the product molecules, the outline of the mapped group would highlight where important subgroups occur.

To correctly implement this idea, the problem of subgraph isomorphism needed to be solved efficiently due to the large number of matches that needed to be performed. The code base used for this project did not include a subgraph isomorphism routine and therefore one had to be implemented from scratch. The approaches mentioned below were all methods implemented to efficiently solve the problem of subgraph isomorphism. Unfortunately, due to the efficiency of the implementations, many of the methods attempted were not viable candidates.

Most subgraph isomorphism algorithms are generally efficient and only become intractable with specific graph structures. These algorithms do come with some graph packages, but as stated earlier the one chosen for this project did not come with a packaged algorithm. The methods implemented all showed signs of intractability

10.1 Brute force

Since the time and space limitations had yet to be tested, a simple brute force approach was implemented (16t, 2002) to test whether it was necessary to worry about the space and time constraints for this problem. Two sample molecules were created, one with 13 atoms and the other with 19 atoms. The brute force method was then applied to the inputs and a result was achieved within 7 minutes of computation.

Unfortunately, this was the result of performing just one subgraph isomorphism. Since the matching had to be done on every reaction in the dataset, the only way to get a solution that fit the constraints would be to either use a very small data set (10 reactions) or use much smaller molecules in the reaction. When this method was implemented, the idea was to use a larger dataset of hundreds of reactions to perform the functional group identification which caused the method to be rejected.

Using smaller molecules in the reactions is not practical as the actual functional groups themselves comprise 6-7 atoms each. So to make the method feasible for the problem, molecules that essentially had no information could only be considered, causing the problem to lose its interest value.

10.2 Error tolerant decomposition

An error tolerant method for inexact matching based on decomposing the graphs to be matched and later recombining was implemented (Messmer and Bunke, 1998). This method decomposes the input graph into its smallest elements and the model graph into a set of unique subgraphs. After this has been done, the units of the input graph are combined together taking note of the errors for the partial subgraph match. By recombining larger and larger subgraphs and remembering the errors associated with the recombinations, the best subgraph matching can be found.

The function used to divide the model graph was chosen to split each subgraph into smaller subgraphs approximately half the size of the original subgraph. This was done as a first approximation for testing.

This method was implemented and tested on an input graph of size 12 atoms with a model graph of 18 atoms. The matching was started, but after 7 minutes of computation and seeing that the method had only found the best match for a graph half the size of the original input graph, the method was abandoned as being to time intensive and not a viable method for this problem.

10.3 Canonicalization

The major setback to the previously described attempts is they all are time intensive. A canonicalization approach was implemented (Owolabi, 1988) next in hopes that by canonicalizing the molecules in the dataset, features could be efficiently extracted and searched for.

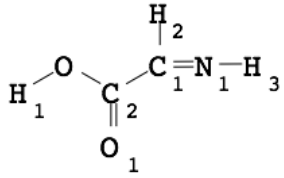

Figure 15: Canonicalization test molecule

Figure 15 shows the molecular structure used to test the canonicalization method. The canonicalization method was tested on this molecule and checked against the results given in (Owolabi, 1988). When the labels associated with the atoms of the same type (such as the carbon atoms in the molecule) were swapped, but the structure was kept the same, the canonicalization matrices did not match. The problem arose because the rules that define how to make the canonicalization matrix still have some inherent

preference for the label. Even though the paper attempts to remove this, it is believed the issue still exists and this causes the error.

Because the method did not correctly canonicalize two structurally equivalent molecules, the method was abandoned.

10.4 Word Segmentation

The previous methods attempted to determine the best match of reactants to products using deterministic techniques. Because the previous methods were all too resource or time intensive, an approach related to machine learning was investigated in hopes that it would alleviate some of the resource requirements.

The problem of functional group identification was thought to be analogous to the problem of word segmentation. Molecules were considered analogous to words and the functional group would be an important segment of the molecule. A method based on word segmentation was implemented with modifications appropriate to the current problem (Goldwater, 2007).

The main assumption with this method was that molecules from different reactions containing the same set of functional groups were able to be identified. Since the functional groups need to be found in all reactions, the functional group should be seen very often. The idea was to find the correct segmentation, based on probabilities, that correctly segments the molecule into functional group and non-functional group segments. A description of the method follows:

A set was created containing products of the separate reactions that shared the same set of functional groups. At this point, the functional groups were not known, it was just known that the products shared the same set of them. Using this set for information, another set S_0 was created by randomly segmenting each of the products in the source set; S_0 is the resulting set of segmentations. In clearer terms, S_0 is defined as follows:

r_x = reaction x reactant having functional group y
$S_s = \{r_1, r_2, r_3, \ldots\}$
seg_x = segmentation of reactant x in S_s
$s_x = \{seg_1, seg_2, seg_3, \ldots\}$
$S_0 = \{s_1, s_2, s_3, \ldots\}$

Beginning with S_0, the probability of each segment was calculated. The probability for a segment belonging to the segmentation s_1 is given by:

$$P_{orig}(seg_1) = \frac{\sum_{i=1}^{|S_0|} count(seg_1 \epsilon S_0[i])}{\sum_{i=1}^{|S_0|} |S_0[i]|}$$

P_{orig} was stored for each segment in S_0 and corresponds to the probability associated with that particular segment. For each pair of segments in s_1 neighboring each other in the unsegmented molecule, the probability of a new segment created by merging the two was calculated. If this new probability is greater than the sum of the individual P_{orig} values for the two segments, in the next iteration of the method, the pair of segments will be replaced by the newly merged segment.

The idea behind this is that the two separate segments could be parts of an overall group. If so, then by joining them the probability of finding the now larger segment would be increased and a larger portion of the actual group, be it functional or not, would be covered. Segments belonging to different groups, when joined together, were thought to show a decrease in the merged probability as the functional group is the only guaranteed group to exist across reactions.

The merging process was continually iterated until no changes in the segmentations were made.

This method failed because it suffered from a problem of clustering. Since the rules only allow for merges and not for the re-splitting of segments, segmentation boundaries can only disappear and are unable to reposition. Neighboring segments can either stay separate as they originally were, or join together to form a larger segment. This resulted in large segments encompassing the original segment used to seed this method.

Another downside to this method is that since molecules are generally connected by carbon atoms, the method had difficulties in determining that carbon atoms connecting the functional group to the less important segments of the molecule are not important. This resulted in the final segmentations surrounding whatever the simplest molecule being examined was.

References

16th International Conference on Pattern Recognition. 2002. *A New Algorithm for Inexact Graph Matching*, volume IV.

Eric V. Anslyn and Dennis A. Dougherty. 2006. *Modern Physical Organic Chemistry*. University Science Books.

Jonathan H. Chen and Pierre Baldi. 2009. No electron left behind: A rule-based expert system to predict chemical reactions and reaction mechanisms. *J. Chem. Inf. Model.*, 49:2034–2043.

William Lingran Chen. 2006. Chemoinformatics: Past, present, and future. *J. Chem. Inf. Model.*, 46(6):2230–2255.

Conference On Information Technology Education. 2007. *Integrating bioinformatics, distributed data management, and distributed computing for applied training in high performance computing*.

Sharon J. Goldwater. 2007. *Nonparametric Bayesian Models of Lexical Acquisition*. Ph.D. thesis, Brown University.

W. L. Jorgensen. 1998. Boss - biochemical and organic simulation system. *Encyclopedia of Computational Chemistry*, 5:3281–3285.

Bruno T. Messmer and Horst Bunke. 1998. A new algorithm for error-tolerant subgraph isomorphism detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(5):493–504, May.

O. Owolabi. 1988. An efficient graph approach to matching chemical structures. *J. Chem. Inf. Comput. Sci.*, 28(4):221–226, February.

Matthew P. Repasky, Jayaraman Chandrasekhar, and William L. Jorgensen. 2002. Pddg/pm3 and pddg/mndo: Improved semiempirical methods. *J. Comput. Chem.*, 23(16):1601–22.

Matthew H. Todd. 2005. Computer-aided organic synthesis. *Chem. Soc. Rev.*, 34:247–266.

K. Peter C. Vollhardt and Neil E. Schore. 2007. *Organic Chemistry: Structure and Function*. W. H. Freeman and Company, 5th edition.